DOCUMENT RESUME

ED 357 085 TM 019 864

AUTHOR Cohen, Allan S.; Kim, Seock-Ho
TITLE A Comparison of Equating Methods under the Graded Response Model.
PUB DATE Apr 93
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chi Square; Comparative Analysis; Computer Simulation; *Equated Scores; *Equations (Mathematics); *Item Response Theory; *Mathematical Models; Research Methodology; Sample Size; *Test Items; Test Length
IDENTIFIERS Calibration; *Graded Response Model; Mean (Statistics); Sigma Methods; Test Characteristic Curve; Test Equivalence

ABSTRACT

Equating tests from different calibrations under item response theory (IRT) requires calculation of the slope and intercept of the appropriate linear transformation. Two methods have been proposed recently for equating graded response items under IRT, a test characteristic curve method and a minimum chi-square method. These two methods are compared with three mean and sigma methods using computer simulations. Ten- and 30-item tests were simulated for 300 and 1,000 examinees. Results under these simulated conditions indicate that recovery is good for all conditions. Recovery is slightly better for the long test and the large sample, but differences among all simulated conditions are quite small. Essentially no differences are observed among the linking methods. One could feel relatively comfortable using any of the five equating methods when ability and item location distributions are well-matched. The simplest equating method is the B. H. Loyd and H. D. Hoover (LH) mean and sigma method (1980). The minimum chi-square has some advantage in ease of use over the test characteristic curve method, but both are more complicated than the LH method. Eight tables present analysis results. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
SEOCK-HO KIM
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED357085

A Comparison of Equating Methods Under the Graded Response Model

Allan S. Cohen and Seock–Ho Kim

University of Wisconsin-Madison

March 19, 1993

Running Head: EQUATING IN THE GRADED RESPONSE MODEL

Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA, April, 1993.

BEST COPY AVAILABLE

TM019864

Abstract

Equating tests from different calibrations under Item Response Theory (IRT) requires calculation of the slope and intercept of the appropriate linear transformation. Two methods have been proposed recently for equating of graded response items under IRT, a test characteristic curve method and a minimum chi-square method. In the present study, we provide a comparison using simulated data sets between these two methods and three mean and sigma methods.

Index terms: equating, graded response model, item response theory.

A Comparison of Equating Methods Under the Graded Response Model

The metrics yielded by current item response theory (IRT) estimation algorithms from separate calibrations for the same items are unique up to a linear transformation. This means that, to equate tests which have been calibrated separately, it is necessary to determine the slope and intercept of the linear equation which yields the appropriate transformation. In the present paper, we compare results from five methods for determining these two transformation coefficients for Samejima's (1969) graded response model.

Three general classes of equating methods have been described for the dichotomous IRT model: characteristic curve methods, minimum chi-square methods, and mean and sigma methods. Characteristic curve methods (cf. Divgi, 1980; Haebara, 1980; Stocking & Lord, 1983) make use of the information available from both the item discrimination and item difficulty parameters. This class of methods is specifically designed to obtain the slope and intercept coefficients by minimizing some measure of the difference between the test characteristic curves estimated in each sample. The Stocking and Lord procedure obtains the two equating coefficients by minimizing a quadratic loss function based on differences in true scores yielded by the two test calibrations. Baker (1992) extended this procedure to the graded response model. The minimum chi-square method proposed by Divgi (1985) for dichotomously scored items, which uses estimates of both item discrimination and difficulty parameters as well as their standard errors, is computationally simpler than the Stocking and Lord procedure. Kim and Cohen (in press) have extended the minimum chi-square method to the graded response model.

Several methods, generally known as mean and sigma methods, have been proposed that rely on the distributions of item difficulty and discrimination estimates

(cf. Bejar & Wingersky, 1981; Cook, Eignor, & Hutton, 1979; Linn, Levine, Hastings, & Wardrop, 1980, 1981; Marco, 1977; Vale, 1986). Mean and Sigma methods are presently only described for the dichotomous model. Comparisons of linking results on dichotomous items suggest that for large samples and long tests, few differences exist among the weighted mean and sigma method of Linn et al. (1980, 1981), the method by Stocking and Lord (1983), and Divgi's (1985) minimum chi- square method (Kim & Cohen, 1992). In the present study, we describe three variations of mean and sigma methods for Samejima's (1969) graded response model and compare the results to those obtained using the methods by Baker (1992) and Kim and Cohen (in press).

Equating Under IRT. Lord (1980) has shown that, under IRT, the relationship of the metric between any two calibrations of the same items from different groups in the same population is linear. Thus, when the estimates from the second calibration are to be transformed to the metric of the first, the transformed estimates of item discrimination and item difficulty parameters of item j for the dichotomously scored, two-parameter IRT model are given by

$$a^*_{j2} = a_{j2}/A \tag{1}$$

and

$$b^*_{j2} = Ab_{j2} + B, \tag{2}$$

where * indicates a transformed value, the subscript 2 refers to the calibration from the second group, A is the slope coefficient, and B is the intercept coefficient.

The value of the transformed ability estimate of person i can be expressed as

$$\theta^*_{i2} = A\theta_{i2} + B. \tag{3}$$

The task of equating the two metrics is to find the appropriate equating coefficients A and B.

Many different equating situations exist (cf. Vale, 1986). In this paper, we consider only that situation for which a set of common items is administered to two groups of examinees.

Samejima's Graded Response Model. Under Samejima's graded response model (Samejima, 1969), an item possesses m_j ordered categories and the examinee is permitted to select only one. Item parameters are estimated under the graded response model via the use of the $m_j - 1$ boundary characteristic curves (BCCs). Each of the BCCs represents the cumulative probablity of selecting response categories greater than the category of interest (Samejima, 1969). The BCCs for item j are characterized by an item discrimination parameter a_j and the $m_j - 1$ location parameters b_{jk}.

The BCC in logistic form can be defined as

$$\hat{P}_{jk}(\theta_i) = [1 + \exp\{-a_j(\theta_i - b_{jk})\}]^{-1} \tag{4}$$

In the case of the metric from the second group being equated to that of the first, the transformation for the graded response model can be obtained via

$$a^*_{j2} = a_{j2}/A \tag{5}$$

and

$$b^*_{jk2} = Ab_{jk2} + B. \tag{6}$$

Samejima (1969) defines the operational characteristic curve (OCC) which shows the probability of selecting a category. The OCC can be obtained from the boundary curves as

$$P_{jk}(\theta_i) = \begin{cases} 1 - \hat{P}_{j1}(\theta_i) & \text{when } k = 1 \\ \hat{P}_{j(m_j-1)}(\theta_i) & \text{when } k = m_j \\ \hat{P}_{j(k-1)}(\theta_i) - \hat{P}_{jk}(\theta_i) & \text{otherwise.} \end{cases} \tag{7}$$

Equating Methods for the Graded Response Model

Mean and Sigma Equating Methods for the Graded Response Model.

Three mean and sigma methods are described in this section for the graded response model: a mean and sigma method (MS) (Marco, 1977), the Loyd and Hoover (1980) method (LH), and a weighted mean and sigma method (WMS) by Linn et al. (1980, 1981).

For the MS method, it is assumed that the location parameter from the first group, $b_{jk1}(j = 1,\ldots,n; k = 1,\ldots,m_j - 1)$, and from the second group, b_{jk2}, are linearly related as

$$b_{jk1} = Ab_{jk2} + B \tag{8}$$

and hence,

$$a_{j1} = a_{j2}/A. \tag{9}$$

The MS equating coefficients A and B are obtained from the following relationships:

$$\bar{b}_1 = A\,\bar{b}_2 + B \tag{10}$$

where $\bar{b}_1$ and $\bar{b}_2$ are the means of the b_{jk1}s and b_{jk2}s, respectively, and

$$S(b_1) = A\,S(b_2) \tag{11}$$

where $S(b_1)$ and $S(b_2)$ are the standard deviations of b_{jk1}s and b_{jk2}s, respectively.

Thus, we can obtain A and B as

$$A = S(b_1)/S(b_2) \tag{12}$$

and

$$B = \bar{b}_1 - A\,\bar{b}_2. \tag{13}$$

Loyd and Hoover (1980) used the ratio of item discrimination parameter estimates from the two calibrations to obtain the A coefficient for their LH method. The LH method was originally used under the Rasch model. Baker and Al-Karni (1991) used the LH method to equate metrics for the three-parameter IRT model. Since we know $a_{j1} = a_{j2}/A$, the A coefficient for the LH method for the graded response model can be obtained as

$$A = \bar{a}_2/\bar{a}_1 \tag{14}$$

and the B coefficient as

$$B = \bar{b}_1 - A\bar{b}_2 \tag{15}$$

where $\bar{a}_2$ and $\bar{a}_1$ are the means of the a_{j1}s and a_{j2}s, respectively.

An important problem with the MS and LH methods is that poorly estimated item difficulties can have a detrimental effect on the values of the A and B coefficients. To overcome this problem, Linn et al. (1980, 1981) modified the MS procedure to include a weighting of item difficulty estimates by the inverse of the larger of the squared standard errors (see also Stocking & Lord, 1983). Stocking and Lord (1983) further scaled this weight by the sum of weights across all items. For the graded response model, the scaled weight for the location parameter estimate for item j and category k, w_{jk}, is defined as

$$w_{jk} = \frac{[\max\{SE(b_{jk1}), SE(b_{jk2})\}]^2}{\sum_{j=1}^{n}\sum_{k=1}^{m_j-1}[\max\{SE(b_{jk1}), SE(b_{jk2})\}]^2}. \tag{16}$$

The weighted estimates of the location parameters are obtained as

$$b^w_{jk1} = w_{jk}b_{jk1} \tag{17}$$

and

$$b_{jk2}^{w} = w_{jk} b_{jk2}. \tag{18}$$

Then, from the relationship

$$\overline{b}_1^{w} = A\overline{b}_2^{w} + B \tag{19}$$

and

$$S(b_1^{w}) = AS(b_2^{w}), \tag{20}$$

the coefficients A and B are obtained as

$$A = S(b_1^{w})/S(b_2^{w}) \tag{21}$$

and

$$B = \overline{b}_1^{w} - A\overline{b}_2^{w}. \tag{22}$$

Two additional refinements to the weighted mean and sigma method, of possible interest, although not treated in the present study, are the robust mean and sigma method described by Bejar and Wingerskky (1981) and the iterative mean and sigma method by Stocking and Lord (1983). The objective of these methods is to further decrease the impact of deviant item location parameter estimates on the linking transformation using biweights described by Mosteller and Tukey (1977).

Test Characteristic Curve Method for Graded Response Model. Baker (1992) extended the test characteristic curve method of Stocking and Lord (1983) to the graded response model. Baker's technique for obtaining the two equating coefficients was based on the minimization of the quadratic loss function

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_{i1} - T_{i2}^{*})^2, \tag{23}$$

where N is an arbitrary number of points along the first ability metric, T_{i1} and T_{i2}^{*} are the true scores for the first and second groups, respectively, defined as

$$T_{i1} = \sum_{j=1}^{n}\sum_{k=1}^{m_j} u_{jk} P_{jk1}(\theta_i) \tag{24}$$

and

$$T_{i2}^{*} = \sum_{j=1}^{n} \sum_{k=1}^{m_j} u_{jk} P_{jk2}^{*}(\theta_i), \tag{25}$$

where u_{jk} is the weight allocated to response category k for item j. Typically, although not necessarily, this weight is the same as the integer index of the category.

The task is to find the values of A and B which minimize the quadratic loss function in Equation (23). In the present study, the characteristic curve method for the graded response model was used as implemented in the computer program EQUATE2 (Baker, 1993).

Minimum Chi-Square for Graded Response Model. Kim and Cohen (in press) extended the minimum chi-square method of Divgi (1985) to the graded response model. The method is based on minimization of the quadratic function

$$\chi^2 = \sum_{j=1}^{n} \chi_{jm_j}^2 = \sum_{j=1}^{n} \underline{\xi}_{jm_j}' \underline{\Sigma}_{jm_j}^{-1} \underline{\xi}_{jm_j}, \tag{26}$$

where

$$\underline{\xi}_{jm_j}' = \underline{\xi}_{jm_j1} - \underline{\xi}_{jm_j2}^{*}, \tag{27}$$

$$\underline{\xi}_{jm_j1} = (a_{j1}, b_{j11}, \ldots, b_{jk1}, \ldots, b_{j(m_j-1)1})', \tag{28}$$

$$\underline{\xi}_{jm_j2}^{*} = (a_{j2}^{*}, b_{j12}^{*}, \ldots, b_{jk2}^{*}, \ldots, b_{j(m_j-1)2}^{*})', \tag{29}$$

and

$$\underline{\Sigma}_{jm_j} = \underline{\Sigma}_{jm_j1} + \underline{\Sigma}_{jm_j2}^{*}, \tag{30}$$

where $\underline{\Sigma}_{jm_j1}$ is the estimated variance-covariance matrix of $\underline{\xi}_{jm_j1}$ and $\underline{\Sigma}_{jm_j2}^{*}$ is the transformed estimated variance-covariance matrix of $\underline{\xi}_{jm_j2}^{*}$. The equating coefficients A and B are found by minimizing this χ^2 differentiating with respect to A and B.

Methods

Data Generation. Data for this study were generated for two test lengths, 10 and 30 items, and two sample sizes, 300 and 1,000 examinees, using the computer program GENIRV (Baker, 1986). The two factors, test length and sample size, were completely crossed to yield four conditions. All items had five categories. Each test was replicated five times by changing the random number seed. Generating parameters for the underlying ability and item difficulty distributions were both normal (0, 1). The underlying item discrimination parameters were generated uniformly over the interval from 1.0 to 2.0. All replication data sets for each of the test lengths had the same set of underlying parameters.

Item Parameter Estimation. Marginal maximum likelihood item parameter estimates were obtained via the computer program MULTILOG (Thissen, 1991). Estimates from each replication were transformed to the metric of the generated data sets using each of the five equating methods. Since the equating task is that of a recovery study, the theoretical values for the linear equating coefficients are known apriori and are $A = 1.0$ and $B = 0.0$.

Results

In this study, the parameter estimates were first transformed to the underlying metric using each of the five equating methods. This yielded five different A and B coefficients for each data set. Next, the recovery of the underlying parameters was evaluated using root mean square differences (RMSDs) between the transformed estimates and the underlying parameters. The smaller the RMSDs, the better the equating method. In addition, correlations between the estimates and the generating parameters were also computed. (Note: Correlations are scale-free meaning that equating is not required.)

Equating Coefficients. Equating coefficients obtained from each of the five

equating methods are given in Tables 1 and 2 for each replication for 300 examinee samples for the 10- and 30- item tests, respectively. Results for the large sample, 1,000 examinee conditions are given in Tables 3 and 4 for the 10- and 30- item tests, respectively.

Insert Tables 1, 2, 3, and 4 about here

Across all data sets, differences in A coefficients were quite small, occasionally arising in the second decimal place but more often in the third or fourth. Differences of this magnitude are essentially zero. In the small sample condition with 300 examinees, differences among A values tended to be very small and not meaningfully different from 1, the theoretically expected value for a recovery study. A values were basically the same for all five equating methods in all four test length by sample size conditions. Differences which did occur were primarily in the second through fourth decimal places and, consequently, were essentially zero. There was a tendency for A values to differ less from 1.0 for the longer 30-item test but none of these differences was greater than .05. No consistent differences were observed among the five equating methods.

Differences in B coefficients also were very small, some ocurring in the second decimal place but more in the third or fourth. As noted for the A coefficients, differences of this magnitude are essentially zero. All of the B coefficients were essentially zero, the theoretically expected value for this recovery study. There was a slight tendency for B values to be closer to zero for the large sample and longer test condition. No consistent differences were observed among the five equating methods.

Recovery of Underlying Parameters. Recovery of the underlying parameters with each method was evaluated with root mean square differences (RMSD) between the transformed estimates and the generating parameters. RMSDs for the 300

examinee samples are given in Tables 5 and 6 for the 10- and 30-item tests, respectively. RMSDs for the 1,000 examinee samples are given in Tables 7 and 8 for the 10- and 30-item tests, respectively. Mean values for RMSDs for both equating coefficients are given for the five equating methods at the bottom of each table. Correlations between estimates and the generating parameters are also given in these tables.

Insert Tables 5, 6, 7, and 8 about here

Recovery of discrimination parameters was good in all data sets. Correlations between estimated discrimination and generating parameters ranged from .731 to .954 in the 300 examinee samples and from .912 to .961 in the 1,000 examinee samples. All correlations indicate good recovery. Mean RMSD values ranged from .1231 to .1494 for the small sample conditions and .0789 to .0903 in the large sample conditions. In addition, smaller RMSDs were observed within each test length for the large sample conditions. The RMSDs for discrimination also indicate good recovery. No differences in recovery were observed among equating methods.

Recovery of location parameters was good under each of the conditions simulated. Correlations with underlying parameters were nearly perfect, ranging from .992 to .999. RMSDs in the 300 examinee test conditions were relatively small (average RMSDs ranged from .1274 to .1449) but were about twice as large as those in the 1,000 examinee conditions (average RMSDs ranged from .0631 to .0744). Values of RMSDs indicated excellent recovery of location parameters. No differences were observed among the five equating methods.

RMSDs showed very slight differences among individual data sets within each of the test length by sample size conditions. For average discrimination or location

parameters, however, no meaningful differences were found among the five equating methods under any of the simulated conditions. What differences were observed were so small (essentially the only differences that were observed were in the second through fourth decimal places) as to be essentially non-existent.

Discussion

The comparability of IRT item parameter estimates across different tests measuring the same underlying trait is an important matter for test developers and researchers since all decisions about examinees are derived from these estimates. Efforts to reduce errors in transformation of estimates obtained in different groups are important concerns. In the present paper, we compared five methods for linking item parameter estimates for graded response models. These five methods are among the more commonly used for transforming item and ability parameter estimates from one metric to another. The comparisons were based on measures of similarity to the generating parameters of the item parameter estimates obtained following transformation via each of the methods to the underlying metric.

Differences in equating coefficients were quite small under all sample size by test length conditions. In the small sample conditions, there was a slight tendency for A and B coefficients to be closer to the theoretically expected values for the 30-item tests. These differences, however, occurred only in the second or third decimal places and, as such, were essentially non-existent. In the large sample conditions, similar lack of deviations from values of 1.0 for A and 0 for B were found.

Results under the conditions simulated indicated that recovery was good for all conditions. Recovery was slightly better for the long test and the large sample conditions but differences among all the simulated conditions actually were quite small. Further, essentially no differences were observed among linking methods. These results are consistent with previous research in that, when the underlying

ability and item difficulty distributions match, estimation of location parameters is optimal and of discrimination parameters tends to be generally good. Recovery of underlying parameters under such conditions also tends to be very good so that differences among equating methods should be quite minimal.

One of the equating methods compared in this study, the minimum chi-square method, required use of the off-diagonal covariance terms for each item. Unfortunately, currently available computer programs do not provide values of these off-diagonal terms so they were not available for the present study. The chi-square (or the quadratic function) that was minimized was obtained based only on the diagonal terms of the variance-covariance matrix. Thus, Σ_{jm_j} in Equation 30 is a diagonal matrix. The resulting statistic, and the one used in the present study, is related to Pearson's (1926) coefficient of racial likeness (CRL). It has been found to be highly correlated with the Mahalanobis D^2 (i.e., $\chi^2_{jm_j}$) in Equation 26 and recommended as a replacement for D^2 because of its computational ease (Gower, 1972; Mardia, 1977; Penrose, 1954).

Finally, given the results of this study, one should feel relatively comfortable using any of the five equating methods when ability and item location distributions are well-matched. That is, when item parameters are estimated under optimal conditions such as used in the present study, little if any real differences appear to be present among these equating methods. Additional research on situations in which item parameters are less well-estimated would be important in further developing our understanding of the effectiveness of each of these equating methods. Under the present conditions, however, the results do not indicate any reason for selecting one method over the other. Neither is there any theoretical rationale for selection one method over the other. The simplest method to use is clearly the LH method. The minimum chi-square method has some advantage in ease of implementation over the test characteristic

curve method but both methods are far more computationally intensive than the LH method.

Graded response models are particularly appropriate for constructed response item formats such as found in many types of performance tests. Development and comparison of the procedures for equating graded response items as was done in this study should provide some useful information toward solving some of the equating problems present in performance and constructed response types of tests.

References

Baker, F. B. (1986). *GENIRV: A program to generate item response vectors* [computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87-96.

Baker, F. B. (1993). *EQUATE2: Computer program for equating two metrics in item response theory* [computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Baker, F. B., & Al-Karni, A. A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147-162.

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81-8). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Hutten, L. R. (1979, April). *Considerations in the application of latent trait theory to objectives-based criterion-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Divgi, D. R. (1980). *Evaluation of scales for multilevel test batteries.* Paper presented at the meting of the American Educational Research Association, Boston, April, 1980.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 77-90.

Gower, J. C. (1972). Measures of taxonomic distance and their analysis. In J. S. Weiner & J. Huizinga (Eds.), *The assessment of population affinities in man* (pp. 1-24). London, England: Oxford University Press.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144-149.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51-66.

Kim, S. H., & Cohen, A. S. (in press). A minimum chi-square method for equating tests under the graded response model. *Applied Psychological Measurement.*

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension* (Technical Report No. 163). Urbana: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159-173.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

Mardia, K. V. (1977). Mahalanobis distance and angles. In P. R. Krishnaiah (Ed.), *Multivariate analysis–IV: Proceedings of the Fourth International Symposium on Multivariate Analysis* (pp. 495-511). Amsterdam, The Netherlands: North-Holland.

Mosteller, F., & Tukey, J.W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison–Wesley.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika, 18*, 105-117.

Penrose, L. S. (1954). Distance, size, and shape. *Annals of Eugenics, 18*, 337-343.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychomteric Monographs, 17.*

Stocking, M., & Lord, F. M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement, 7*, 207-210.

Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 6.0): [computer program]. Chicago, IL: Scientific Software.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333-344.

TABLE 1

Equating Coefficients A and B for 300-Examinee-10-Item Data Set

Rep.[a]	Equating Coefficient	Method				
		LH	MS	WMS	MCS	TCC
1st	*A*	.9197	.9257	.9592	.9357	.9220
	B	.0659	.0665	.0021	.0829	.0725
2nd	*A*	1.0232	.9731	1.0132	1.0078	1.0009
	B	.0554	.0496	.0008	.0328	.0477
3rd	*A*	.9599	.9190	.9244	.9465	.9352
	B	−.0252	−.0257	−.0008	−.0364	−.0294
4th	*A*	.9970	.9910	1.0322	1.0124	.9878
	B	−.0270	−.0264	−.0011	−.0427	−.0265
5th	*A*	.9778	.9631	.9901	.9819	.9706
	B	.0032	.0020	−.0003	−.0125	−.0075

[a]Replication.

TABLE 2

Equating Coefficients A and B for 300-Examinee-30-Item Data Set

Rep.[a]	Equating Coefficient	Method				
		LH	MS	WMS	MCS	TCC
1st	*A*	.9943	.9811	1.0173	1.0064	.9861
	B	.0198	.0191	.0001	.0147	.0218
2nd	*A*	1.0336	.9857	1.0402	1.0256	1.0105
	B	−.0454	−.0449	−.0004	−.0468	−.0494
3rd	*A*	1.0305	1.0082	1.0332	1.0280	1.0174
	B	−.0362	−.0362	−.0004	−.0499	−.0399
4th	*A*	1.0561	1.0437	1.0606	1.0569	1.0461
	B	−.0028	−.0030	.0001	.0069	−.0028
5th	*A*	1.0185	.9928	1.0279	1.0185	1.0018
	B	−.0147	−.0151	−.0001	−.0067	−.0122

[a]Replication.

TABLE 3

Equating Coefficients A and B for 1000-Examinee-10-Item Data Sets

Rep.[a]	Equating Coefficient	Method				
		LH	MS	WMS	MCS	TCC
1st	*A*	.9695	.9580	.9575	.9628	.9627
	B	−.0295	−.0302	−.0006	−.0260	−.0286
2nd	*A*	.9836	.9654	.9792	.9779	.9681
	B	−.0146	−.0153	−.0002	−.0085	−.0160
3rd	*A*	.9504	.9424	.9600	.9511	.9470
	B	−.0266	−.0266	−.0006	−.0261	−.0243
4th	*A*	.9785	1.0025	1.0079	.9961	.9917
	B	.0013	.0026	−.0001	−.0049	.0022
5th	*A*	.9780	.9706	.9734	.9722	.9658
	B	.0103	.0108	.0002	.0074	.0096

[a]Replication.

TABLE 4

Equating Coefficients A and B for 1000-Examinee-30-Item Data Sets

Rep.[1]	Equating Coefficient	Method				
		LH	MS	WMS	MCS	TCC
1st	*A*	1.0165	1.0103	1.0253	1.0172	1.0155
	B	−.0165	−.0165	−.0002	−.0208	−.0203
2nd	*A*	1.0160	1.0135	1.0272	1.0205	1.0143
	B	.0086	.0080	.0000	.0023	.0067
3rd	*A*	.9965	.9858	1.0091	.9955	.9905
	B	−.0105	−.0104	−.0001	−.0129	−.0113
4th	*A*	.9854	.9818	.9932	.9882	.9827
	B	−.0228	−.0226	−.0002	−.0182	−.0218
5th	*A*	.9935	.9879	1.0083	.9960	.9902
	B	−.0036	−.0037	−.0001	−.0088	−.0069

[1]Replication.

TABLE 5

Root Mean Squared Differences and Correlation for 300-Examinee-10-Item Data Sets

Rep.[a]	Parameter	Method					Corr.
		LH	MS	WMS	MCS	TCC	
1st	Discrimination	.0752	.0751	.0958	.0780	.0750	.954
	Location	.1115	.1116	.1389	.1139	.1117	.995
2nd	Discrimination	.1602	.1833	.1617	.1632	.1658	.731
	Location	.1606	.1459	.1647	.1552	.1515	.992
3rd	Discrimination	.1196	.1453	.1930	.1241	.1309	.903
	Location	.1497	.1379	.1470	.1444	.1406	.993
4th	Discrimination	.1339	.1353	.1388	.1334	.1364	.881
	Location	.1276	.1270	.1401	.1318	.1268	.994
5th	Discrimination	.1268	.1305	.1272	.1266	.1281	.847
	Location	.1300	.1279	.1336	.1319	.1290	.994
Average	Discrimination	.1231	.1339	.1433	.1251	.1272	.863
	Location	.1359	.1301	.1449	.1354	.1319	.994

[a]Replication.

TABLE 6

Root Mean Squared Differences and Correlation for 300-Examinee-30-Item Data Sets

Rep.[a]	Parameter	Method					Corr.
		LH	MS	WMS	MCS	TCC	
1st	Discrimination	.1234	.1273	.1246	.1228	.1254	.891
	Location	.1307	.1288	.1396	.1340	.1292	.994
2nd	Discrimination	.1550	.1772	.1545	.1564	.1616	.777
	Location	.1566	.1426	.1663	.1529	.1474	.992
3rd	Discrimination	.1296	.1374	.1292	.1300	.1331	.876
	Location	.1136	.1094	.1200	.1137	.1105	.996
4th	Discrimination	.1303	.1327	.1300	.1302	.1321	.837
	Location	.1235	.1220	.1244	.1240	.1221	.995
5th	Discrimination	.1623	.1725	.1609	.1623	.1678	.851
	Location	.1394	.1344	.1435	.1396	.1354	.993
Average	Discrimination	.1401	.1494	.1398	.1403	.1440	.847
	Location	.1328	.1274	.1388	.1328	.1289	.994

[a]Replication.

TABLE 7

Root Mean Squared Differences and Correlation for 1000-Examinee-10-Item Data Sets

Rep.[a]	Parameter	Method					Corr.
		LH	MS	WMS	MCS	TCC	
1st	Discrimination	.0653	.0689	.0692	.0668	.0668	.959
	Location	.0444	.0420	.0513	.0427	.0425	.999
2nd	Discrimination	.0956	.1038	.0968	.0972	.1020	.955
	Location	.0855	.0820	.0855	.0841	.0822	.997
3rd	Discrimination	.0966	.0985	.0967	.0965	.0972	.923
	Location	.0674	.0664	.0749	.0675	.0669	.998
4th	Discrimination	.0895	.0972	.1007	.0938	.0920	.912
	Location	.0708	.0661	.0667	.0667	.0669	.998
5th	Discrimination	.0830	.0831	.0827	.0828	.0844	.961
	Location	.0592	.0591	.0603	.0593	.0593	.999
Average	Discrimination	.0860	.0903	.0892	.0874	.0885	.942
	Location	.0655	.0631	.0677	.0641	.0636	.998

[a]Replication.

TABLE 8

Root Mean Squared Differences and Correlation for 1000-Examinee-30-Item Data Sets

Rep.[a]	Parameter	Method					Corr.
		LH	MS	WMS	MCS	TCC	
1st	Discrimination	.0829	.0835	.0839	.0829	.0829	.923
	Location	.0695	.0690	.0734	.0698	.0695	.998
2nd	Discrimination	.0792	.0795	.0798	.0791	.0794	.943
	Location	.0713	.0712	.0739	.0721	.0712	.998
3rd	Discrimination	.0776	.0801	.0792	.0777	.0786	.937
	Location	.0734	.0719	.0784	.0732	.0723	.998
4th	Discrimination	.0772	.0777	.0775	.0771	.0776	.943
	Location	.0771	.0708	.0759	.0716	.0709	.998
5th	Discrimination	.0779	.0788	.0800	.0778	.0783	.939
	Location	.0676	.0671	.0702	.0683	.0673	.998
Average	Discrimination	.0790	.0799	.0801	.0781	.0794	.937
	Location	.0655	.0631	.0677	.0641	.0636	.998

[a]Replication.